http://dx.doi.org/10.14236/ewic/EVA2024.5

What Is New in AI – More of the Same or an Artificial General Intelligence Breakthrough?

Susan Hazan
Digital Heritage, Israel
Maagal Beit Hamidrash 22
Jerusalem, Israel
Susan.hazan@gmail.com

1. INTRODUCTION

Throughout 2023 the AI sector continued to command significant public interest when the launch of Google's new large language model (LLM) Gemini was offered to developers and enterprise customers at the end of the year and made headlines for what seemed like its impressive multimodal performance in processing images, videos, and audio. Even though Google later admitted to the widely published criticism that the promotional video was 'fabricated or tweaked' the launch caused quite a stir (Edwards 2023). The demo in the video (2024) appeared to show Gemini identifying objects and relationships in visual data, challenging the user to entertaining games, while solving self-improvised scenarios. At the same time, the highly popular image-generating models available in the public sector were still enjoying expediential growth throughout the year with new and impressive versions such as DALL·E 3 and Midjourney v.6 released to the public. Both models proved to be incrementally better than their previous versions and both continued to dazzle and excite with new functionalities and variations. At the same time, Open AI released its beta version of Sora, a much-hyped but rather underwhelming video generator. According to Open AI, Today, Sora has been offered to red teamers to assess critical areas for harm or risks, as well as granting access to some visual artists, designers, and filmmakers to gain feedback on how to advance the model to be most helpful for creative professionals. 2023 had been a very prolific year for AI developers and the public was more than happy not only to experiment with these systems but to actively integrate their functionalities into their working and creative lives.

The AI sector presented numerous opportunities for users to sign up for a whole range of tantalizing platforms – either for payment or for free while at the same companies were already eying their next goal. The next critical landmark was hotly anticipated in research centres and the commercial sector that saw artificial general intelligence (AGI) as their ultimate goal. There had been many red herrings dropped into the public discourse concerning whether AI had already reached the point where LLMs such as GPT-4 models had attained the level where users identified sentience in their interaction with these systems or something approaching human consciousness. See for example when Microsoft's OpenAI-powered Bing chatbot spontaneously changed its name from Sidney to Venon and had been making serious waves by making up horror stories, gaslighting users, passive-aggressively admitting defeat, and generally being extremely unstable in incredibly bizarre ways (Hazan 2024). This apparently spooked the research community but most of the calls of alarm seemed to be aimed at hyping their very expensive products, perhaps in a bid to create a demand for them. AGI is mainly theoretical, but frenetic development is underway to create platforms with human-like intelligence including the much-coveted abilities such as self-teaching, emergent agency, and autonomy, and even the suggestion of self-replication. This capability was suspected with Google's Bard spontaneously learning to speak Bengali, a language which it had not been specifically trained for. The claim by Sundar Pichai, CEO of Google's parent company Alphabet, was later challenged with the explanation that AI had already been trained in Bengali with millions of tokens. Reaching AGI still seems to be mainly elusive but the race to create an autonomous system that surpasses human capabilities is more than tantalizing.

© Hazan. Published by BCS Learning and Development Ltd. Proceedings of EVA London 2024, UK

2. TOWARDS AGI

So how can we know with confidence that we are witness to an AGI event or entity, the moment when software performs as well or better than humans on a wide range of cognitive activities or shows the capability to self-teach and even self-replicate? For decades the world has been pondering singularity – the moment AI becomes the dominant form of intelligence on Earth as robots or computer software effectively takes control of our planet. First coined by John von Neumann in 1958, the term “Singularity” describes a hypothesis that developments in science – especially in the fields of informatics, nanotechnology, and biology will ultimately lead to the invention of an AGI. Ray Kurzweil famously discussed how artificial machines with natural intelligence would be safe or not and based on scientific, philosophical, and theological arguments suggested how human-level AGI could evolve as early as around 2029, and reach singularity in 2045. While there are many ways where companies and research centres are negotiating their distinct pathways towards realizing AGI in addition to the public and secret government programs around the world this paper suggests just three such pathways out of the many.

3. AGENCY

Already freely available and highly popular are the intelligent AI-driven agents, intelligent agents that can be programmed to support a range of tasks either as a simple logical progression or in the more sophisticated model-based system that is focused on a complex goal. An AI agent can be instructed to act on your behalf. For example, you could instruct your agent on the occasion of your five-year-old birthday to help with all the arrangements. You could ask it to find a suitable hall, pick out the perfect children’s entertainer for the party, perhaps a magician or local band. After making all your choices you then prompt the system to act on these plans and book the room, the magician, make the catering arrangements, and even purchase the gifts on your behalf. Although these are innocuous actions once a less benign agent is granted agency to a system to act on their behalf in the real world, one that perhaps had been mal programmed with bugs in the system or even biases that crept in during the training the results could just as easily be nefarious. At what point does our AI actor break out of its prescribed boundaries and what happens when things get out of control? These scenarios could be very unnerving.

4. AI GETS AN INNER MONOLOGUE

One of the notable limitations of LMs is their inability to contextualize the flow of ideas causing them to ‘hallucinate’, an innocuous term that can be otherwise described as confabulation or delusion. Not only does this remind us that AIs are untrustworthy, but their blatant predilection for lying becomes a major obstacle for a system that attempts sentiency. Quiet Self-Taught Reasoner, Quiet-STaR, supports a process that resembles a human's inner monologue and supports a reasoning process that enrichens outputs, allowing users of the model to be able to select which response is most accurate like the reasoning that is implicit in almost all written text. LMs learn to generate rationales at each token to explain the future text, thereby improving their predictions. The process represents the text (as think), mixing the future-text predictions with and without rationales (talk), and then learning to generate better rationales using REINFORCE (learn) Zelikman et al P.2) and delivers ideas that are closer to natural text, connecting the meaning hidden between the lines – the missing context.

5. SELF-REPLICATION

In a New York Times podcast with Ezra Klein (2024) Anthropic CEO Dario Amodei raising truly existential questions by suggesting AI models might reach autonomy levels whereby AI entities can self-replicate and survive, a possibility he suggests that could happen as early as 2025-2028. This poses formidable governance challenges and hair-raising security risks. This is but one progression that is not beyond impossibility in the future AGI landscape and one that is truly chilling. The Computer Gone Rogue has played out in our collective memories since the spine-chilling movie 2001: A Space Odyssey (1968) in the famous scene where Hal 9000 computer refuses to obey an order from Bowman (Keir Dullea) by simply responding in monotone, "I'm sorry Dave, I'm afraid I can't do that." Although there are fortunes to be made by the company that first cracks the AGI code – one can only hope that whoever gets there first treads with extreme caution and leads the way with responsibility and care.

6. REFERENCES

Al-Sibon (2024) Scientists Gave AI an _Inner Monologue_ and Something Fascinating Happened. https://futurism.com/the-byte/ai-inner-monologue

DALL·E 3 (2024) https://openai.com/dall-e-3

Edwards, B. (2023) Google's best Gemini AI demo video was fabricated, Ars Technica, https://arstechnica.com/information-technology/2023/12/google-admits-it-fudged-a-gemini-ai-demo-video-which-critics-say-misled-viewers/

Google (2024) The capabilities of multimodal AI | Gemini Demo. https://youtu.be/UIZAiXYceBI?si=VnVZnGdVhhqP3gNJ

Hazan, S. (2024) The Promise of Pastiche: How ai Systems Pervade the Cultural Heritage Sector, in Arts and Computational Culture, Tula Giannini and Jonathan P. Bowen, Springer.

Klein, E. (2004) What if Dario Amodei Is Right About A.I.?, NYT Podcast. https://www.nytimes.com/2024/04/12/podcasts/transcript-ezra-klein-interviews-dario-amodei.html

Kurzweil, R. (2005) The Singularity Is Near: When Humans Transcend Biology. Penguin Books, New York.

von Neumann, J. (1958) *The Computer and the Brain*. New Haven: Yale University Press, pp.66-82

Sora (2024) https://openai.com/sora

Stable Video (2024) https://www.stablevideo.com/

Zelikman E. et al. (2024) Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking. https://arxiv.org/abs/2403.09629